

Holland & Knight LLP
701 Brickell Avenue
Suite 3000
Miami, Florida 33131
Telephone: (305) 789-7773

Application for United States Letters Patent

filed on behalf of

Applicant: Arun Ramagopal
For: System and Method for Identifying
a Network Resource

Attorney Docket: 91353-1

PATENT

SYSTEM AND METHOD FOR IDENTIFYING A NETWORK RESOURCE

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] Not Applicable.

5

STATEMENT REGARDING FEDERALLY SPONSORED-RESEARCH OR DEVELOPMENT

[0002] Not Applicable.

10 **INCORPORATION BY REFERENCE OF MATERIAL SUBMITTED ON A COMPACT DISC**

[0003] Not Applicable.

FIELD OF THE INVENTION

15 [0004] The invention disclosed broadly relates to the field of information technologies and more particularly relates to the field of firewalls and transmission of network resources.

BACKGROUND OF THE INVENTION

20 [0005] HTTP is the most common protocol in use for web browsing and file downloads. It is a TCP-based protocol and thus data packets are sent and received in an orderly manner by both the client and server. Data packets using this protocol comprise two parts: header information and data. An HTTP proxy server is a common network node that decodes the HTTP protocol, and is currently one of several network

PATENT

gateway devices used by network administrators to limit access by nodes in an intranet or local area network (LAN) to the Internet. For example, pornography sites, email sites such as Hotmail, and sports sites are commonly blocked at corporation network gateway devices. This is generally done through an HTTP proxy server installed at
5 the LAN, by eliminating certain IP addresses from the LAN's local DNS server, or by adding IP-based restrictions at any other node. These network gateway devices scan the incoming request for the destination domain name or IP address. If the field matches a set of known Internet locations (IP addresses or domain names) then the request is blocked. The set of Internet locations is normally maintained by hand by
10 the network administrators who installed the network gateway device. However, blocking unwanted resources from the Internet is a challenging task. Much of this difficulty is due to the fact that the information needing to be scanned can be a combination of the header and data part of the packet, packets are considered stateless, and the specific data sections (offsets) to scan are constantly changing due to new and
15 evolving Internet-enabled programs and DNS entries.

SUMMARY OF THE INVENTION

[0006] Briefly according to the invention, a method comprises steps of routing network communication comprising one or more packets, each packet comprising
20 bytes structured according to the Internet Protocol (IP); gathering and storing unordered packets in memory in order to effectively scan UDP-based protocols; scanning the bytes of one or more packets to extract identifying information relating to the network resource; comparing the extracted identifying information to a set of identifying information stored in a database; using a central server farm that constantly
25 finds the identifying information to be filtered and updates each database; and

PATENT

providing a message indicating that the extracted information matches at least one entry in the database when the comparison is positive.

BRIEF DESCRIPTION OF THE DRAWINGS

- 5 [0007] FIG. 1 is an illustration of a network comprising a system according to the present invention.
- [0008] FIGs. 2 - 4 show various configuration of local area networking using the invention.
- 10 [0009] FIG. 5 is a high level flow chart illustrating a method according to the invention.
- [0010] FIG. 6 illustrates a system for identifying network resources.
- [0011] FIGs. 7a - 7b show a flowchart illustrating a detailed method according to an embodiment of the invention.
- 15 [0012] FIG. 8 shows an HTTP GET Method request where structure information is only in the header section.
- [0013] FIG.9 shows an HTTP POST Method request structure where information is in both the header and data sections.
- [0014] FIG. 10 shows the response from a server to an HTTP request.
- 20 [0015] FIG. 11 shows a Peer to Peer request using Fasttrack communication and a hash code.
- [0016] FIG.12 shows a Peer to Peer request using Fasttrack communication and a filename.
- [0017] FIG. 13 shows the response from a server using Fasttrack communication to a Peer to Peer request.

PATENT

[0018] FIG. 14 shows a Peer to Peer request using Gnutella communication and a filename.

[0019] FIG. 15 shows a response from a server to a Peer to Peer request using Gnutella communication.

5 [0020] FIG. 16 shows a retrieved resource using a File Transfer Protocol.

DETAILED DESCRIPTION

[0021] Referring to FIG. 1, there is shown a block diagram of a local area network 100 comprising network gateway devices (NGD) 102 according to an embodiment of the invention. In the embodiment shown in FIG. 1, the LAN 100 comprises a plurality of NGDs 102 (represented by the two shown), each serving a set of client personal computer units 101. The NGDs 102 protect their clients 101 from access to undesired resources by routing packets either received from the WAN 110 or from clients 101 and comparing identifying information such as metadata about network resources in the packets with identifying information stored in a database 103. The database 103 is shown as a shared resource but the network 100 can also be implemented with a database 103 embedded in each NGD 102 so that it can be accessed directly through its API. In any case each database is regularly updated. When the comparison is positive (i.e., a match is found), the NGD 102 provides a message indicating the match. The message can either be displayed as a warning that the content may be inappropriate or misappropriated or to trigger one of various ways of filtering (filtering includes tracking and blocking) the access.

[0022] "Identifying information" is information found in the received stream of packets that is useful for deciding whether to provide access to the network resource. The database 103 is updated to include identifying information relating to

resources to which access by clients is to be controlled. The database 103 can be either shared as shown in FIG. 1 or can be integrated into each of the NGDs 102. In either case, a communication process is in place to update the identifying information for all databases in the system such that the databases operate in a real time manner.

5 The identifying information can be any information that can be extracted or derived from the packets, being transferred throughout the networks 100 and 110 that can be used to identify a resource comprising one or more of the packets.

[0023] In a preferred embodiment, the set of metadata changes for the application being used. The first scanning step of NGD102 is to determine the application being used by the client. In its current embodiment, applications supported are as follows: 1) Web browsers, 2) the Peer 2 Peer programs based on the Fasttrack and Gnutella protocols, specifically Kazaa, Morpheus, Grokster, and their clones, 3) FTP programs, and 4) specialized SMTP junkmail programs such as WorldCast that allow users to run a local SMTP server and bypass their ISP's SMTP

10 server.

[0024] For Web browsers, there are two scanning algorithms that take place along with two sets of metadata. The first scanning algorithm bases its decision on the following metadata obtained from the data packet stream and contained in the database 103: 1) IP address, 2) port, 3) path, 4) resource or file name. As an example,

20 in the following theoretical scenario an HTTP client sends the following request:

1 GET /illegalfiles/IllegalResource.zip HTTP/1.0

2 HOST: www.illegalhost.com

3 [BLANK_LINE][END_OF_STREAM]

[0025] The NGD 102 understands the HTTP application-level protocol, and thus extracts the following information: 1) the IP address based on NGD 102's DNS lookup of the domain name, or directly if the IP address is contained in the client's request, 2) if the port is not contained in the request, the default HTTP port, 80, is

25

used, 3) the path contained in Line 1 above, and 4) the resource as identified by Line 1 above. Since illegalhost.com is an example, 127.0.0.1 will be the theoretical IP address found after domain name resolution. Thus, the extracted information is as follows: 1) 127.0.0.1, 2) 80, 3) illegalfiles, 4) IllegalResource.zip. In this
5 embodiment, this is all the information needed by NGD 102 to effectively block very specific network resources for this HTTP request method.

[0026] If it is determined by the NGD 102 that further scanning is needed because the resource contains an HTML form or processing is needed for the query string, then additional metadata is extracted and examined from the same data packet
10 stream. This additional metadata is as follows: 5) HTML form name-value pairs. In its current embodiment, this information is stored in the same table as described above in the Database with column 5 optional. As an example, in the following scenario the HTTP client sends the following request:

```
1 POST /forms/webform.html HTTP/1.0  
15 2 HOST: www.illegalhost.com  
3 [BLANK_LINE]  
4 resource=IllegalResource.zip&user=username  
5 [BLANK_LINE][END_OF_STREAM]
```

[0027] The HTTP post method sends an unlimited amount of HTML form data
20 after the blank line so that it is considered the data portion of HTTP communication and does not have any size restrictions. This allows HTML forms to contain fields that are very large. In contrast, if a webpage contains an HTML form that contains small fields, it is very common to use the GET method. The following HTTP request has the same purpose as above, but uses the GET method and embeds the form values
25 in the Query String:

```
1 GET /forms/webform.html?resource=IllegalResource.zip&user=username  
HTTP/1.0
```

2 HOST: www.illegalhost.com

3 [BLANK_LINE]][END_OF_STREAM]

In these two scenarios, the form values can be used to request a resource and must be understood by NGD 102 in order to effectively block the transmission. Thus, the
5 following information is extracted: 1) 127.0.0.1, 2) 80, 3) forms, 4) webform.html, 5) resource=IllegalResource.zip. It ignores the username of the form since in this theoretical case the CSF (central server farm) has decided this field is not necessary for NGD 102 to determine the resource. If this information is found in the Database, the network resource transmission is ended.

10 [0028] The LAN 100 supports a packet-switched protocol and is connected to a wide area network 110 (such as the Internet) by means of a conventional firewall 108. The LAN 100 can also comprise a conventional load balancer 106 disposed between the NGDs 102 and the firewall 108 and a conventional router 104 disposed between the load balancer and the NGDs 102.

15 [0029] FIG. 2 illustrates an embodiment of the invention wherein the NGDs 102 are each connected to the firewall 108 by means of the load balancer 106.

[0030] FIG. 3 illustrates an embodiment of the invention wherein the router 104 includes an NGD 102 and the router is disposed between the firewall 108 and the client computers 101.

20 [0031] FIG. 4 illustrates an embodiment of the invention wherein the firewall 108 comprises an NGD 102.

[0032] The network gateway device is preferably an open standard generic application proxy server that combines firewall technologies and application-level resource filtering techniques. It preferably complies with the most common proxy
25 server standards used, such as SOCKS versions 4 and 5. It is preferably implemented with the fastest and most reliable cross-platform programming language available, such as Java 1.4.2. The NGD 102 can be used to do any of the following:

[0033] The NGD 102 can warn users that it appears they are downloading illegal material. This is a service that ISPs and schools can provide to their users.

[0034] The NGD 102 can block specific network resources such as application, music, or movie files that appear to be pirated versions of the material. It is at the network manager's discretion to allow full blocking or to allow illegal downloads to continue with the warning described above. The NGD 102 supports both types of behavior, although blocking is the preferred solution.

[0035] The NGD 102 can block specific programs based on their application-level protocols from being transmitted within that LAN. These protocols can use either TCP (Transmission Control Protocol) or UDP (User Datagram Protocol). For instance, if an ISP (Internet Service Provider) decides that the Kazaa program should not be run on the LAN, the NGD 102 can be configured to support this behavior.

[0036] The NGD 102 can also limit access to external SMTP hosts by only allowing users to make direct TCP connections to specified SMTP servers that the LAN can monitor. This prevents users from sending junk emails from that LAN.

[0037] The NGD 102 can also prevent external users from downloading illegal material from users within the LAN.

[0038] The NGD 102 provides generic support for any IP-based application-level protocol which uses TCP or UDP. In its current embodiment, this is done by conforming to the SOCKS protocol and providing application-level resource-filtering algorithms when necessary. The application-level protocols supported are taken from current versions of TCP-based and UDP-based applications, such as Peer2Peer, HTTP, FTP, and IRC programs. The NGD 102 preferably uses the data that is sent with these programs to analyze the network communication between any client and server. Based on this stream of data packets, The NGD 102 can stop the communication at any point or warn users of activity not supported by their LAN.

[0039] A core feature of the NGD 102 is the implementation of a self-updating and real time database. Each database 103 table maps directly to metadata used by application-level protocols in order for NGD 102 to block specific network resources that these protocols are being used to request. There are tables for the HTTP, FTP, 5 Fasttrack, and Gnutella protocols. In its preferred embodiment, NGD 102 does not use the database 103 for limiting access to SMTP hosts, but instead uses a configuration setting.

[0040] The tables in the current embodiment of the database 103 contain the 10 following columns:
HTTP: "IP address", "port", "path", "resource name", "priority"
FTP: "IP address", "path", "resource name", "priority"
Fasttrack: "Fasttrack Hash Code", "priority"
Gnutella: "SHA1 Hash Code", "priority"
15 P2P-Alternate: "IP address", "port", "identity-key", "resource name", "priority"

[0041] In its preferred embodiment, the Database 103 synchronizes its data with the Central Server Farm in a near real-time manner by listening on a specified port. Whenever a Database 103 starts, even if embedded within an NGD 102, it 20 contacts the CSF and registers its currently configured IP address and port. Thus, the CSF uses its list of Database 103s to send a message signifying either a new entry in or a removal from the Database 103. Database 103s may also request a full synchronization or update at any time by contacting the CSF. In a default installation of the preferred embodiment, a full synchronization happens daily at 12AM in order to 25 maintain each Database 103's data integrity. This allows for the following unique benefits: (1) The protected material is always current. (2) Wrongfully blocked material can be removed in a near real-time fashion. (3) A daily log from each NGD

is sent to a data warehouse containing only the metadata which caused a blocked request. This data contains the same information in the Database tables described above, and is used only to determine the NGD's effectiveness. For instance, in the case of a Fasttrack network resource transmission block, the following information is
5 logged: "Fasttrack Hash Code".

[0042] The NGD 102 will actively filter against the following five protocols:

- 1) HTTP;
- 2) FTP;
- 3) SMTP;
- 10 4) Fasttrack; and
- 5) Gnutella

However, the NGD 102 can easily be adapted to prevent or warn of access to resources in network modes using different protocols.

[0043] The NGD 102 is preferably a SOCKS versions 4 and 5 implementation
15 as described above that also understands the hypertext transfer protocol and other common application-level protocols. Because of this combination of technologies and its unique scanning algorithms, the NGD 102 supports the following additional services that a traditional HTTP proxy server does not:

- 1) Scanning additional header fields besides the host field;
- 20 2) Identifying and scanning additional protocols that use nonstandard HTTP headers known as HTTP extensions;
- 3) Scanning the data portion of HTTP communication, that is, the bytes occurring after the first blank line as per the HTTP specification;
- 4) Using the information contained in the database in order to filter requests. This
25 database is self-updating, and thus does not allow tampering or the involvement of a network administrator.

[0044] The NGD 102 can also interpret HTTP form data based on the specific webpage where the form exists.

FTP

5 [0045] FTP is one of the oldest TCP protocols. A client uses one connection in order to maintain a session with a server. This communication is also analyzed by NGD 102. Many hackers use public FTP sites to host illegal files for a short period of time. These sites are known as 0-day sites, and are referred to as such because on the 1st day an accessible site is discovered (day 0) its utility rating is 100%. The owner of
10 the site does not yet know it is being used for illegal purposes, and not many users know the IP address. By day 10, the usefulness of the site is said to be at 1/1000th of the utility level of day 0. At this point, many users have discovered the IP address and the site's owner may be notified of the security breach. When this happens, the hackers remove the IP address from their lists.

15 [0046] Hackers are in constant search of public web or FTP sites in which to store their files. Many of these servers are in other countries and thus are impossible to shut down by United States laws. Yahoo! Groups (TM) is another common public storage facility for hackers. Specific groups are created simply to distribute files.

[0047] Because of the near real-time Database 103, a system using the
20 invention can actively protect against 0-day web and FTP sites. Only specific file requests are blocked, and so public access to the FTP site is never restricted by the NGD 102. Similarly, Yahoo! Groups and similar web sites are not blocked as a whole, but rather only specific files stored on these sites are.

25 Fasttrack and Gnutella Peer2Peer protocols

[0048] Both Fasttrack and Gnutella use an extended version of HTTP as the primary transport protocol for downloads. This provides reliability and stability for

large file downloads. Although UDP and HTTPS are used for communication with and discovery of peers on the network, all programs currently use HTTP as the download protocol.

5 [0049] This fact allows NGD 102 to block or warn against downloads by matching the file signatures found in the request against the Database 103. HTTP is not encrypted and thus NGD 102 is free to analyze any portion of the network communication.

[0050] The notion of a hash code is very important to all Fasttrack and Gnutella clients. Fasttrack defines the "Fasttrack Hash Code", while Gnutella has the
10 "SHA1." The use of hash codes is an evolution of previous Peer 2 Peer protocols, and allows a client to easily identify any file among hundreds of millions, or billions, of files. It is analogous to a fingerprint in that each hash code is a unique file signature. Several websites exist to catalog hash codes. These files have been verified to be the real working version, and not a decoy or corrupted file. These are the three
15 most popular websites that perform this service:

<http://www.verifieddownloads.com/>

<http://www.fasttrackmovies.com/>

<http://www.fasttrackcentral.com/>

20 [0051] In addition to providing a unique identity, hash codes allows for one client to download from an arbitrary number of servers. With a broadband connection, a user can typically download the same file from 16 different users at the same time. The client then puts the file back together. This ability is incredibly powerful and, at the current time, is only possible due to hash codes.

25 [0052] The blocking of hash code-based Peer 2 Peer protocols is effective because all Peer 2 Peer programs that NGD 102 currently supports use extended

HTTP for the download protocol. In the case of popular Fasttrack client Kazaa, a theoretical request structure is as follows:

```
1 GET /.hash=d0633f1bfdd0fde48cf351ef8c541b67567426dd HTTP/1.1
2 Host: 123.52.193.31:1214
5 3 User-Agent: KazaaClient Jul 20 2003 23:25:14
4 X-Kazaa-Username: logn
5 X-Kazaa-Network: KaZaA
6 X-Kazaa-IP: 213.77.151.176:2647
7 X-Kazaa-SupernodeIP: 206.158.106.142:1715
10 8 Connection: close
9 X-Kazaa-XferId: 11312345
10 X-Kazaa-XferUid: ytCcDgo+3sTohNI2+1Y2jYkCY6NwCA==
```

[0053] In the case of popular Gnutella client Morpheus, a theoretical request structure is as follows:

```
15 1 GET http://81.65.32.7:6346/uri-
res/N2R?urn:sha1:F3HBAWBPQWOS5G5GBCDBPYDMG5NZIA2P HTTP/1.1
2 Host: 81.65.32.7:6346
3 User-Agent: Morpheus 3.3.0.24 (GnucDNA 0.9.2.6)
20 4 Listen-IP: 206.170.247.13:13484
5 Connection: Keep-Alive
6 Proxy-Connection: close
7 Range: bytes=104144-524287
8 X-Queue: 0.1
25 9 X-Gnutella-Content-URN:
urn:sha1:F3HBAWBPQWOS5G5GBCDBPYDMG5NZIA2P
```

[0054] In both cases, a hash code is extracted as per the application-level's protocol and matched against Database 103. Currently, this hash code is embedded into Line 1 for both Kazaa and Morpheus, but the NGD 102 can extract it from other sections in the same manner.

[0055] If a protocol does not use hash codes, it is very difficult to download from two or more peers from the same time. For these protocols, the NGD 102 uses the near real-time information constantly being gathered by the CSF and sent to each NGD 102, and basis its blocking decision on the unique resource request structure the

protocol uses. For instance, Fasttrack and Gnutella define an alternate download method that is also used as the primary download protocol for dozens of less popular Peer 2 Peer programs to interoperate. In this scenario, a user can generally only download any given resource from one single peer at a time. This alternate protocol
5 does not include the hash code as part of the client request but rather appends a unique number to the beginning of the requested resource name.

[0056] The NGD 102 handles these protocols by relying on the CSF to constantly monitor the peers on the supported non-hash code Peer 2 Peer networks, download resources from the peers and match them against the CSF's data warehouse,
10 and send one packet of information to update the Database 103 if the resource is considered illegal by the CSF. In the following scenario where the CSF is monitoring the Grokster Peer 2 Peer network, the CSF is constantly searching for the term "Michael Jackson Thriller", downloading the resource from any peer which is hosting this file according to Grokster's search algorithm, and verifying it to be illegal against
15 the CSF data warehouse. As an example, the CSF finds this resource on a Grokster peer whose IP address is 163.118.98.30 and is listening on port 3504, and updates the P2P-Alternate Database 103 table with the following information: 1) 163.118.98.30, 2) 3504, 3) 14160, 4) Michael Jackson – Thriller.mp3, 1. This information is found because the CSF uses Grokster itself to download the material and thus has access to
20 its protocol. This example would use the following request structure:

1 GET /14160/Michael%20Jackson%20-%20Thriller.mp3 HTTP/1.1

2 Host: 163.118.98.30:3504

3 UserAgent: KazaaClient May 28 2002 14:48:42

4 X-Kazaa-Username: login

25 5 X-Kazaa-Network: Grokster

6 X-Kazaa-IP: 127.0.0.1:0

7 X-Kazaa-SupernodeIP: 67.161.65.106:2167

8 Connection: close

9 X-Kazaa-XferId: 1610030

[0057] After being updated with this new resource's identifying information
5 by the CSF, NGD 102 can extract the same information and end the transmission if a
match against Database 103 is found.

UDP

10 [0058] UDP is used to send individual packets from one machine to another.
The NGD 102 routes UDP packets but may not filter them. It performs this
functionality to comply with the SOCKS version 5 protocol. The NGD 102 must
always support UDP since it may someday be used as a download protocol. Since
UDP is a stateless protocol and there is no guarantee for the arrival or ordering of the
15 packets, the NGD 102 will hold the packets in memory and interpret these packets by
re-ordering them according to their application-level protocol. For instance, in a
typical client/server communication where UDP is used, some packets may or may not
arrive, and if they do arrive it is not understood implicitly by the IP-layer what order
they should be processed. This must be done explicitly by the client and server. As an
20 example, if the client is sending three UDP packets to a server and order and reliability
is to be maintained, the client must specify the order in one or more bytes of the UDP
packet. If the NGD 102 determines that the UDP packet is being sent by an
application-level protocol that is must filter, then it finds the bytes specifying order,
holds all three packets in memory, re-orders the bytes, and filters this in-memory data
25 packet stream as described above. Thus, if the resource identifying information is
anywhere in the three packets, or a combination of the three packets, the NGD 102
will be able to find the necessary metadata.

[0059] It should be noted that this functionality is not used by the NGD 102 in its preferred embodiment as all current NGD 102 supported application-level protocols use TCP. It is programmatically difficult to ensure reliable client/server communication using UDP. Thus TCP has become the *de facto* standard for IP communication and is used by the vast majority of clients and servers. It is believed that UDP will someday be used to try and circumvent NGD 102.

SMTP

[0060] SMTP is the Internet's primary mail protocol. A spammer (sender of junk email) generally makes direct connections to external SMTP servers using DNS Mail Exchange routing. This bypasses the ISP's internal SMTP server, and thus the user is free to mask their identity and hide their actions from the ISP.

[0061] When NGD 102 detects a TCP connection to an SMTP server, it can stop this connection. If an ISP chooses to use this functionality, it is required to set known SMTP servers which their users are allowed to use. All other SMTP server communication will be stopped.

Instant Messaging (TM)

[0062] Instant Messaging (TM) programs use their own protocols. The Internet Engineering Task Force is currently standardizing one protocol for all programs to use.

[0063] Therefore, while there has been described what is presently considered to be the preferred embodiment, it will be understood by those skilled in the art that other modifications can be made within the spirit of the invention.

What is claimed is: